The supplementary materials is organized as follows. In Appendices A, B and C, we provide detailed proofs of the theoretical results in the paper. In Appendix D, we provide additional figures for the experiments described in Section 5.

A. Proof of Theorem 1

In this appendix we prove the minimax bound of Theorem 1. The result is obtained by combining the following two lower bounds:

Theorem 3 (Lower bound 1). *For each problem instance such that $\mathbb{E}_\mu[\rho^2\sigma^2] < \infty$, we have*

$$R_n(\pi;\lambda,\mu,\sigma,R_{\max}) \geq \frac{\mathbb{E}_\mu[\rho^2\sigma^2]}{32en}\left[1 - \frac{\mathbb{E}_\mu\left[\rho^2\sigma^2\mathbf{1}\left(\rho\sigma^2 > R_{\max}\sqrt{n\mathbb{E}_\mu[\rho^2\sigma^2]/2}\right)\right]}{\mathbb{E}_\mu[\rho^2\sigma^2]}\right]^2 .$$

Theorem 4 (Lower bound 2). *Assume that $\mathbb{E}_\mu[\rho^2 R_{\max}^2] < \infty$, and we are given $\gamma \in [0,1]$ and $\delta \in (0,1]$. Write $\xi \coloneqq \xi_\gamma$ and $\gamma' \coloneqq \max\{\gamma,\delta\}$. Then there exist functions $\hat{R}(x,a)$ and $\hat{\rho}(x,a)$ such that*

$$\hat{R}^2(x,a) \leq R_{\max}^2(x,a) \leq (1+\delta)\hat{R}^2(x,a) \ , \qquad \hat{\rho}^2(x,a) \leq \rho^2(x,a) \leq (1+\delta)\hat{\rho}^2(x,a)$$

and the following lower bound holds:

$$R_n(\pi;\lambda,\mu,\sigma,R_{\max}) \geq \frac{\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]}{32en}\left[1 - \frac{\mathbb{E}_\mu\left[\xi\hat{\rho}^2\hat{R}^2\mathbf{1}\left(\xi\hat{\rho}\hat{R} > \sqrt{n\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]/16}\right)\right]}{\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]}\right]^2 - \gamma'\log(5/\gamma')(1+\delta)\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2] \ .$$

The reason for introducing γ' in Theorem 4 is to allow $\gamma = 0$, which is an important special case of the theorem. Otherwise, we could just assume $0 < \delta \leq \gamma$. The first bound captures the intrinsic difficulty due to the variance of reward, and is present even in a vanilla multi-armed bandit problem without contexts. The second result shows the additional dependence on $R_{\max}^2$, even when $\sigma \equiv 0$, whenever the distribution λ is not too degenerate, and captures the additional difficulty of the contextual bandit problem. We next show how these two lower bounds yield Theorem 1 and then return to their proofs.

Proof of Theorem 1. Throughout the theorem we write $\xi \coloneqq \xi_\gamma$. We begin by simplifying the two lower bounds. Assume that Assumption 1 holds with ϵ. This also means that $\mathbb{E}_\mu[\xi(\rho R_{\max})^{2+\epsilon}]$ is finite as well as $\mathbb{E}_\mu[\xi(\rho R_{\max})^2]$ is finite and either both of them are zero or both of them are non-zero. Similarly, $\mathbb{E}_\mu[(\rho\sigma)^{2+\epsilon}]$ and $\mathbb{E}_\mu[(\rho\sigma)^2]$ are both finite and either both of them are zero or both of them are non-zero, so C_γ is a finite constant. Let $p = 1+\epsilon/2$ and $q = 1+2/\epsilon$, i.e., $1/p + 1/q = 1$. Further, let $\hat{R}$ and $\hat{\rho}$ be the functions from Theorem 4. Then the definition of C_γ means that

$$\begin{aligned} C_\gamma^{1/(\epsilon q)} = C_\gamma^{1/(2+\epsilon)} &= 2\cdot\max\left\{\frac{\mathbb{E}_\mu\left[\xi(\rho^2R_{\max}^2)^{\frac{2+\epsilon}{2}}\right]^{\frac{2}{2+\epsilon}}}{\mathbb{E}_\mu\left[\xi\rho^2R_{\max}^2\right]}, \frac{\mathbb{E}_\mu\left[(\rho^2\sigma^2)^{\frac{2+\epsilon}{2}}\right]^{\frac{2}{2+\epsilon}}}{\mathbb{E}_\mu\left[\rho^2\sigma^2\right]}\right\} \\ &= 2\cdot\max\left\{\frac{\mathbb{E}_\mu\left[\xi(\rho^2R_{\max}^2)^p\right]^{1/p}}{\mathbb{E}_\mu\left[\xi\rho^2R_{\max}^2\right]}, \frac{\mathbb{E}_\mu\left[(\rho^2\sigma^2)^p\right]^{1/p}}{\mathbb{E}_\mu\left[\rho^2\sigma^2\right]}\right\} \\ &\geq 2\cdot\max\left\{\frac{\mathbb{E}_\mu\left[\xi(\hat{\rho}^2\hat{R}^2)^p\right]^{1/p}}{\mathbb{E}_\mu\left[\xi\rho^2R_{\max}^2\right]}, \frac{\mathbb{E}_\mu\left[(\rho^2\sigma^2)^p\right]^{1/p}}{\mathbb{E}_\mu\left[\rho^2\sigma^2\right]}\right\} \ , \end{aligned} \tag{10}$$

and recall that we assume that

$$n \geq \max\left\{16C_\gamma^{1/\epsilon},\, 2C_\gamma^{2/\epsilon}\mathbb{E}_\mu[\sigma^2/R_{\max}^2]\right\} \ . \tag{11}$$

First, we simplify the correction term in the lower bound of Theorem 3. Using Hölder's inequality and Eq. (10), we have

$$\begin{aligned}
&\mathbb{E}_\mu\left[\rho^2\sigma^2\mathbf{1}\left(\rho\sigma^2 > R_{\max}\sqrt{n\mathbb{E}_\mu[\rho^2\sigma^2]/2}\right)\right] \\
&\le \mathbb{E}_\mu\left[\left(\rho^2\sigma^2\right)^p\right]^{1/p}\cdot\mathbb{P}_\mu\left[\rho\sigma^2 > R_{\max}\sqrt{n\mathbb{E}_\mu[\rho^2\sigma^2]/2}\right]^{1/q} \\
&\le \frac{1}{2}\mathbb{E}_\mu[\rho^2\sigma^2]\cdot C_\gamma^{1/(\epsilon q)}\cdot\mathbb{P}_\mu\left[\rho\sigma^2/R_{\max} > \sqrt{n\mathbb{E}_\mu[\rho^2\sigma^2]/2}\right]^{1/q}.
\end{aligned}$$

We further invoke Markov's inequality, Cauchy-Schwartz inequality, and Eq. (11) in the following three steps to simplify this event as

$$\begin{aligned}
&\le \frac{1}{2}\mathbb{E}_\mu[\rho^2\sigma^2]\cdot C_\gamma^{1/(\epsilon q)}\cdot\left(\frac{\mathbb{E}_\mu\left[\rho\sigma\cdot(\sigma/R_{\max})\right]}{\sqrt{n\mathbb{E}_\mu[\rho^2\sigma^2]/2}}\right)^{1/q} \\
&\le \frac{1}{2}\mathbb{E}_\mu[\rho^2\sigma^2]\cdot C_\gamma^{1/(\epsilon q)}\cdot\left(\frac{\sqrt{\mathbb{E}_\mu[\rho^2\sigma^2]}\cdot\sqrt{\mathbb{E}_\mu[\sigma^2/R_{\max}^2]}}{\sqrt{n\mathbb{E}_\mu[\rho^2\sigma^2]/2}}\right)^{1/q} \\
&= \frac{1}{2}\mathbb{E}_\mu[\rho^2\sigma^2]\cdot\left(C_\gamma^{2/\epsilon}\cdot\frac{2\mathbb{E}_\mu[\sigma^2/R_{\max}^2]}{n}\right)^{1/2q} \le \frac{1}{2}\mathbb{E}_\mu[\rho^2\sigma^2]\ .
\end{aligned} \tag{12}$$

For the correction term in Theorem 4, we similarly have

$$\begin{aligned}
&\mathbb{E}_\mu\left[\xi\hat\rho^2\hat R^2\mathbf{1}\left(\xi\hat\rho\hat R > \sqrt{n\mathbb{E}_\mu[\xi\hat\rho^2\hat R^2]/16}\right)\right] \\
&\le \mathbb{E}_\mu\left[\left(\xi\hat\rho^2\hat R^2\right)^p\right]^{1/p}\cdot\mathbb{P}_\mu\left[\xi\hat\rho\hat R > \sqrt{n\mathbb{E}_\mu[\xi\hat\rho^2\hat R^2]/16}\right]^{1/q} \\
&\le \frac{1}{2}\mathbb{E}_\mu[\xi\rho^2R_{\max}^2]\cdot C_\gamma^{1/(\epsilon q)}\cdot\mathbb{P}_\mu\left[\xi\hat\rho^2\hat R^2 > n\mathbb{E}_\mu[\xi\hat\rho^2\hat R^2]/16\right]^{1/q},
\end{aligned}$$

so that Markov's inequality and Eq. (11) further yield

$$\begin{aligned}
&\le \frac{1}{2}\mathbb{E}_\mu[\xi\rho^2R_{\max}^2]\cdot C_\gamma^{1/(\epsilon q)}\cdot\left(\frac{\mathbb{E}_\mu[\xi\hat\rho^2\hat R^2]}{n\mathbb{E}_\mu[\xi\hat\rho^2\hat R^2]/16}\right)^{1/q} \\
&= \frac{1}{2}\mathbb{E}_\mu[\xi\rho^2R_{\max}^2]\cdot\left(C_\gamma^{1/\epsilon}\cdot\frac{16}{n}\right)^{1/q} \le \frac{1}{2}\mathbb{E}_\mu[\xi\rho^2R_{\max}^2] \le \frac{(1+\delta)^2}{2}\mathbb{E}_\mu[\xi\hat\rho^2\hat R^2]\ .
\end{aligned} \tag{13}$$

Using Eq. (12), the bound of Theorem 3 simplifies as

$$\begin{aligned}
&R_n(\pi;\lambda,\mu,\sigma,R_{\max}) \\
&\ge \frac{\mathbb{E}_\mu[\rho^2\sigma^2]}{32en}\left[1-\frac{\mathbb{E}_\mu\left[\rho^2\sigma^2\mathbf{1}\left(\rho\sigma^2 > R_{\max}\sqrt{n\mathbb{E}_\mu[\rho^2\sigma^2]/2}\right)\right]}{\mathbb{E}_\mu[\rho^2\sigma^2]}\right]^2 \\
&\ge \frac{\mathbb{E}_\mu[\rho^2\sigma^2]}{32en}\left(1-\frac{1}{2}\right)^2 = \frac{\mathbb{E}_\mu[\rho^2\sigma^2]}{128en}\ .
\end{aligned} \tag{14}$$

Similarly, by Eq. (13), Theorem 4 simplifies as

$$
\begin{aligned}
&R_n(\pi;\lambda,\mu,\sigma,R_{\max}) \\
&\quad\geq \frac{\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]}{32en}\left[1-\frac{\mathbb{E}_\mu\left[\xi\hat{\rho}^2\hat{R}^2\mathbf{1}\left(\xi\hat{\rho}\hat{R}>\sqrt{n\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]/16}\right)\right]}{\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]}\right]^2 - \gamma'\log(5/\gamma')(1+\delta)\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2] \\
&\quad\geq \frac{\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]}{32en}\left[1-\frac{(1+\delta)^2}{2}\right]^2 - \gamma'\log(5/\gamma')(1+\delta)\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2] \\
&\quad= \frac{\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2]}{128en}\left(1-2\delta-\delta^2\right)^2 - \gamma'\log(5/\gamma')(1+\delta)\mathbb{E}_\mu[\xi\hat{\rho}^2\hat{R}^2] \\
&\quad\geq \frac{\mathbb{E}_\mu[\xi\rho^2R_{\max}^2]}{128en}\frac{\left(1-2\delta-\delta^2\right)^2}{(1+\delta)^2} - \gamma'\log(5/\gamma')(1+\delta)\mathbb{E}_\mu[\xi\rho^2R_{\max}^2]\ .
\end{aligned}
$$

Since this bound is valid for all $\delta>0$, taking $\delta\to 0$, we obtain

$$
R_n(\pi;\lambda,\mu,\sigma,R_{\max}) \geq \frac{\mathbb{E}_\mu[\xi\rho^2R_{\max}^2]}{128en} - \gamma\log(5/\gamma)\mathbb{E}_\mu[\xi\rho^2R_{\max}^2]\ .
$$

Combining this bound with Eq. (14) yields

$$
\begin{aligned}
&R_n(\pi;\lambda,\mu,\sigma,R_{\max}) \\
&\quad\geq \frac{1}{2}\cdot\frac{\mathbb{E}_\mu[\rho^2\sigma^2]}{128en} + \frac{1}{2}\cdot\frac{\mathbb{E}_\mu[\xi\rho^2R_{\max}^2]}{128en} - \frac{1}{2}\cdot\gamma\log(5/\gamma)\mathbb{E}_\mu[\xi\rho^2R_{\max}^2] \\
&\quad\geq \frac{\mathbb{E}_\mu[\rho^2\sigma^2]}{700n} + \frac{\mathbb{E}_\mu[\xi\rho^2R_{\max}^2]}{700n} - \frac{1}{2}\cdot\gamma\log(5/\gamma)\mathbb{E}_\mu[\xi\rho^2R_{\max}^2] \\
&\quad= \frac{1}{700n}\left[\mathbb{E}_\mu[\rho^2\sigma^2] + \mathbb{E}_\mu[\xi\rho^2R_{\max}^2]\left(1-350n\gamma\log(5/\gamma)\right)\right]\ .
\end{aligned}
$$

□

It remains to prove Theorems 3 and 4. They are both proved by a reduction to hypothesis testing, and invoke Le Cam's argument to lower-bound the error in this testing problem. As in most arguments of this nature, the key contribution lies in the construction of an appropriate testing problem that leads to the desired lower bounds. Before proving the theorems, we recall the basic result of Le Cam which underlies our proofs. We point the reader to the excellent exposition of Lafferty et al. (2008, Section 36.4) on more details about Le Cam's argument.

Theorem 5 (Le Cam's method, Lafferty et al., 2008, Theorem 36.8)**.** *Let $\mathcal{P}$ be a set of distributions, let $X_1,\ldots,X_n$ be an i.i.d. sample from some $P\in\mathcal{P}$, let $\theta(P)$ be any function of $P\in\mathcal{P}$, let $\hat{\theta}(X_1,\ldots,X_n)$ be an estimator, and d be a metric. For any pair $P_0,P_1\in\mathcal{P}$,*

$$
\inf_{\hat{\theta}}\sup_{P\in\mathcal{P}}\mathbb{E}_P[d(\hat{\theta},\theta(P))] \geq \frac{\Delta}{8}e^{-nD_{\mathrm{KL}}(P_0\|P_1)} \tag{15}
$$

where $\Delta = d(\theta(P_0),\theta(P_1))$, and $D_{\mathrm{KL}}(P_0\|P_1)=\int\log(dP_0/dP_1)dP_0$ is the KL-divergence.

While the proofs of the two theorems share a lot of similarities, they have to use reductions to slightly different testing problems given the different mean and variance constraints in the two results. We begin with the proof of Theorem 3, which has a simpler construction.

A.1. Proof of Theorem 3

The basic idea of this proof is to reduce the problem of policy evaluation to that of Gaussian mean estimation where there is a mean associated with each x,a pair. We now describe our construction.

Creating a family of problems Since we aim to show a lower bound on the hardness of policy evaluation in general, it suffices to show a particular family of hard problem instances, such that every estimator requires the stated number of

samples on at least one of the problems in this family. Recall that our minimax setup assumes that π, μ and λ are fixed and the only aspect of the problem which we can design is the conditional reward distribution $D(r \mid x, a)$. For Theorem 3, this choice is further constrained to satisfy $\mathbb{E}[r \mid x, a] \le R_{\max}(x, a)$ and $\mathrm{Var}(r \mid x, a) \le \sigma^2(x, a)$. In order to describe our construction, it will be convenient to define the shorthand $\mathbb{E}[r \mid x, a] = \eta(x, a)$. We will identify a problem in our family with the function $\eta(x, a)$ as that will be the only changing element in our problems. For a chosen η, the policy evaluation question boils down to estimating $v_\eta^\pi = \mathbb{E}[r(x, a)]$, where the contexts x are chosen according to λ, actions are drawn from $\pi(x, a)$ and the reward distribution $D_\eta(r \mid x, a)$ is a normal distribution with mean $\eta(x, a)$ and variance $\sigma^2(x, a)$

$$D_\eta(r \mid x, a) = \mathcal{N}(\eta(x, a),\, \sigma^2(x, a)).$$

Clearly this choice meets the variance constraint by construction, and satisfies the upper bound so long as $\eta(x, a) \le R_{\max}(x, a)$ almost surely. Since the evaluation policy π is fixed throughout, we will drop the superscript and use v_η to denote v_η^π in the remainder of the proofs. With some abuse of notation, we also use $\mathbb{E}_\eta[\cdot]$ to denote expectations where contexts and actions are drawn based on the fixed choices λ and μ corresponding to our data generating distribution, and the rewards drawn from η. We further use P_η to denote this entire joint distribution over (x, a, r) triples.

Given this family of problem instances, it is easy to see that for any pair of η_1, η_2 which are both pointwise upper bounded by $R_{\max}$, we have the lower bound:

$$R_n(\lambda, \pi, \mu, \sigma^2, R_{\max}) \ge \inf_{\hat{v}} \max_{\eta \in \eta_1, \eta_2} \mathbb{E}_\eta\Big[\underbrace{(\hat{v} - v_\eta)^2}_{\ell_\eta(\hat{v})}\Big],$$

where we have introduced the shorthand $\ell_\eta(\hat{v})$ to denote the squared error of $\hat{v}$ to v_η. For a parameter $\epsilon > 0$ to be chosen later, we can further lower bound this risk for a fixed $\hat{v}$ as

$$\begin{aligned} R_n(\hat{v}) &\ge \max_{\eta \in \eta_1, \eta_2} \mathbb{E}_\eta[\ell_\eta(\hat{v})] \ge \max_{\eta \in \eta_1, \eta_2} \epsilon \mathbb{P}_\eta(\ell_\eta \ge \epsilon) \\ &\ge \frac{\epsilon}{2}\Big[\mathbb{P}_{\eta_1}(\ell_{\eta_1}(\hat{v}) \ge \epsilon) + \mathbb{P}_{\eta_2}(\ell_{\eta_2}(\hat{v}) \ge \epsilon)\Big], \end{aligned} \tag{16}$$

where the last inequality lower bounds the maximum by the average. So far we have been working with an estimation problem. We next describe how to reduce this to a hypothesis testing problem.

Reduction to hypothesis testing For turning our estimation problem into a testing problem, the idea is to identify a pair η_1, η_2 such that they are far enough from each other so that any estimator which gets a small estimation loss can essentially identify whether the data generating distribution corresponds to P_{η_1} or P_{η_2}. In order to do this, we take any estimator $\hat{v}$ and identify a corresponding test statistic which maps $\hat{v}$ into one of η_1, η_2. The way to do this is essentially identified in Eq. (16), and we describe it next.

Note that since we are constructing a hypothesis test for a specific pair of distributions P_{η_1} and P_{η_2}, it is reasonable to consider test statistics which have knowledge of η_1 and η_2, and hence the corresponding distributions. Consequently, these tests also know the true policy values v_{η_1} and v_{η_2} and the only uncertainty is which of them gave rise to the observed data samples. Therefore, for any estimator $\hat{v}$, we can a associate a statistic $\phi(\hat{v}) = \mathrm{argmin}_\eta \{\ell_{\eta_1}(\hat{v}), \ell_{\eta_2}(\hat{v})\}$.

Given this hypothesis test, we are interested in its error rate $\mathbb{P}_\eta(\phi(\hat{v}) \ne \eta)$. We first relate the estimation error of $\hat{v}$ to the error rate of the test. Suppose for now that

$$\ell_{\eta_1}(\hat{v}) + \ell_{\eta_1}(\hat{v}) \ge 2\epsilon, \tag{17}$$

so that at least one of the losses is at least ϵ. Suppose that the data comes from η_1. Then if $\ell_{\eta_1}(\hat{v}) < \epsilon$, we know that the test is correct, because by Eq. (17) the other loss is greater than ϵ, and therefore $\phi(\hat{v}) = \eta_1$. This means that the error under η_1 can only occur if $\ell_{\eta_1}(\hat{v}) \ge \epsilon$. Similarly, the error under η_2 can only occur if $\ell_{\eta_2}(\hat{v}) \ge \epsilon$, so the test error can be bounded as

$$\begin{aligned} \max_{\eta \in \eta_1, \eta_2} \mathbb{P}_\eta(\phi(\hat{v}) \ne \eta) &\le \mathbb{P}_{\eta_1}(\phi(\hat{v}) \ne \eta_1) + \mathbb{P}_{\eta_2}(\phi(\hat{v}) \ne \eta_2) \\ &\le \mathbb{P}_{\eta_1}(\ell_{\eta_1}(\hat{v}) \ge \epsilon) + \mathbb{P}_{\eta_2}(\ell_{\eta_2}(\hat{v}) \ge \epsilon) \\ &\le \frac{2}{\epsilon} R_n(\hat{v}), \end{aligned} \tag{18}$$

where the final inequality uses our earlier lower bound in Eq. (16).

To finish connecting our the estimation problem to testing, it remains to establish our earlier supposition (17). Assume for now that η_1 and η_2 are chosen such that

$$(v_{\eta_1} - v_{\eta_2})^2 \geq 4\epsilon. \tag{19}$$

Then an application of the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ yields

$$4\epsilon \leq (v_{\eta_1} - v_{\eta_2})^2 \leq 2(\hat{v} - v_{\eta_1})^2 + 2(\hat{v} - v_{\eta_2})^2 = 2\ell_{\eta_1}(\hat{v}) + 2\ell_{\eta_2}(\hat{v}),$$

which yields the posited bound (16).

Invoking Le Cam's argument So far we have identified a hypothesis testing problem and a test statistic whose error is upper bounded in terms of the minimax risk of our problem. In order to complete the proof, we now place a lower bound on the error of this test statistic. Recall the result of Le Cam (15), which places an upper bound on the attainable error in any testing problem. In our setting, this translates to

$$\max_{\eta \in \eta_1, \eta_2} \mathbb{P}_\eta(\phi(\hat{v}) \neq \eta) \geq \frac{1}{8} e^{-n D_{\mathrm{KL}}(P_{\eta_1} \,||\, P_{\eta_2})}.$$

Since the distribution of the rewards is a spherical Gaussian, the KL-divergence is given by the squared distance between the means, scaled by the variance, that is

$$D_{\mathrm{KL}}(P_{\eta_1} \,||\, P_{\eta_2}) = \mathbb{E}\left[\frac{(\eta_1(x,a) - \eta_2(x,a))^2}{2\sigma^2(x,a)}\right],$$

where we recall that the contexts and actions are drawn from λ and μ respectively. Since we would like the probability of error in the test to be a constant, it suffices to choose η_1 and η_2 such that

$$\mathbb{E}\left[\frac{(\eta_1(x,a) - \eta_2(x,a))^2}{2\sigma^2(x,a)}\right] \leq \frac{1}{n}. \tag{20}$$

Picking the parameters So far, we have not made any concrete choices for η_1 and η_2, apart from some constraints which we have introduced along the way. Note that we have the constraints (19) and (20) which try to ensure that η_1 and η_2 are not too close that an estimator does not have to identify the true parameter, or too far that the testing problem becomes trivial. Additionally, we have the upper and lower bounds of 0 and $R_{\max}$ on η_1 and η_2. In order to reason about these constraints, it is convenient to set $\eta_2 \equiv 0$, and pick $\eta_1(x,a) = \eta_1(x,a) - \eta_2(x,a) = \Delta(x,a)$. We now write all our constraints in terms of Δ.

Note that v_{η_2} is now 0, so that the first constraint (19) is equivalent to

$$v_{\eta_1} = \mathbb{E}_{\eta_1}[\rho(x,a)r(x,a)] = \mathbb{E}_{\Delta}[\rho(x,a)r(x,a)] \geq 2\sqrt{\epsilon},$$

where the importance weighting function ρ is introduced since P_{η_1} is based on choosing actions according to μ and we seek to evaluate π. The second constraint (20) is also straightforward

$$\mathbb{E}\left[\frac{\Delta^2}{2\sigma^2}\right] \leq \frac{1}{n}.$$

Finally, the bound $R_{\max}$ and non-negativity of η_1 and η_2 are enforced by requiring $0 \leq \Delta(x,a) \leq R_{\max}(x,a)$ almost surely.

The minimax lower bound is then obtained by the largest ϵ in the constraint (19) such that the other two constraints can be satisfied. This gives rise to the following variational characterization of the minimax lower bound:

$$\begin{aligned}
\max_{\Delta} \quad & \epsilon \\
\text{such that} \quad & \mathbb{E}_{\Delta}[\rho(x,a)r(x,a)] \geq 2\sqrt{\epsilon}, && (21)\\
& \mathbb{E}\left[\frac{\Delta^2}{2\sigma^2}\right] \leq \frac{1}{n}, && (22)\\
& 0 \leq \Delta(x,a) \leq R_{\max}(x,a). && (23)
\end{aligned}$$

Instead of finding the optimal solution, we just exhibit a feasible setting of Δ here. We set

$$\Delta = \min\left\{\frac{\alpha\sigma^2\rho}{\mathbb{E}_\mu[\rho^2\sigma^2]}, R_{\max}\right\}, \quad \text{where} \quad \alpha = \sqrt{\frac{2\mathbb{E}_\mu[\rho^2\sigma^2]}{n}}. \tag{24}$$

This setting satisfies the bounds (23) by construction. A quick substitution also verifies that the constraint (22) is satisfied. Consequently, it suffices to set ϵ to the value attained in the constraint (21). Substituting the value of Δ in the constraint, we see that

$$\begin{aligned}
\mathbb{E}_\Delta[\rho(x,a)r(x,a)] &= \mathbb{E}_{x\sim\lambda,a\sim\mu}[\rho(x,a)\Delta(x,a)] \\
&\geq \mathbb{E}_{x\sim\lambda,a\sim\mu}\left[\rho\frac{\alpha\sigma^2\rho}{\mathbb{E}_\mu[\rho^2\sigma^2]}\mathbf{1}\Big(\rho\sigma^2\alpha \leq R_{\max}\mathbb{E}_\mu[\rho^2\sigma^2]\Big)\right] \\
&= \alpha\left(1 - \frac{\mathbb{E}_\mu\left[\rho^2\sigma^2\mathbf{1}\big(\rho\sigma^2\alpha > R_{\max}\mathbb{E}_\mu[\rho^2\sigma^2]\big)\right]}{\mathbb{E}_\mu[\rho^2\sigma^2]}\right) \\
&=: 2\sqrt{\epsilon}.
\end{aligned}$$

Putting all the foregoing bounds together, we obtain that for all estimators $\hat{v}$

$$\begin{aligned}
R_n(\hat{v}) &\geq \frac{\epsilon}{2}\cdot\Big(\max_{\eta\in\eta_1,\eta_2}\mathbb{P}_\eta(\phi(\hat{v})\neq\eta)\Big) \\
&\geq \frac{\epsilon}{2}\cdot\frac{1}{8}e^{-nD_{\mathrm{KL}}(P_{\eta_1}\,||\,P_{\eta_2})} \\
&\geq \frac{\epsilon}{2}\cdot\frac{1}{8e} = \frac{\epsilon}{16e} \\
&= \frac{1}{16e}\cdot\frac{\alpha^2}{4}\left(1 - \frac{\mathbb{E}_\mu\left[\rho^2\sigma^2\mathbf{1}\big(\rho\sigma^2 > R_{\max}\mathbb{E}_\mu[\rho^2\sigma^2]/\alpha\big)\right]}{\mathbb{E}_\mu[\rho^2\sigma^2]}\right)^2 \\
&= \frac{\mathbb{E}_\mu[\rho^2\sigma^2]}{32en}\left(1 - \frac{\mathbb{E}_\mu\left[\rho^2\sigma^2\mathbf{1}\big(\rho\sigma^2 > R_{\max}\sqrt{n\mathbb{E}_\mu[\rho^2\sigma^2]/2}\big)\right]}{\mathbb{E}_\mu[\rho^2\sigma^2]}\right)^2.
\end{aligned}$$

A.2. Proof of Theorem 4

We now give the proof of Theorem 4. While it shares a lot of reasoning with the proof of Theorem 3, it has one crucial difference. In Theorem 3, there is a non-trivial noise in the reward function, unlike in Theorem 4. This allowed the proof to work with just two candidate mean-reward functions, since any realization in the data is corrupted with noise. However, in the absence of added noise, the task of mean identification becomes rather trivial: an estimator can just check whether η_1 or η_2 matches the observations exactly.

To prevent such a strategy, we instead construct a richer family of reward functions. Instead of merely two mean rewards, our construction will involve a randomized design of the expected reward function from an appropriate prior distribution. The draw of the mean reward from a prior will essentially generate noise even though any given problem is noiseless. The construction will also highlight the crucial sources of difference between the contextual and multi-armed bandit problems, since the arguments here rely on having access to a rich context distribution, by which we mean distribution that puts non-trivial probability on many contexts. In the absence of this property, the bound of Theorem 4 becomes weaker.

Creating a family of problems Our family of problems will be parametrized by the two reals δ and γ from the statement of the theorem. Our construction begins with a discretization step at the resolution δ, whose goal is to create a countable partition of the set of pairs $\mathcal{X}\times\mathcal{A}$. If sets $\mathcal{X}$ and $\mathcal{A}$ are countable or finite, this step is vacuous, but if the sets of contexts or actions have continuous parts, this step is required.

First, let $\mu(x,a)$ denote the joint probability measure obtained by first drawing $x\sim\lambda$ and then $a\sim\mu(\cdot\,|\,x)$. In Lemma 1, we show that $\mathcal{X}\times\mathcal{A}$ can be split into countably many disjoint sets B_i, $\biguplus_{i\in\mathcal{I}} B_i = \mathcal{X}\times\mathcal{A}$, such that the following conditions are satisfied:

- Each $i \in \mathcal{I}$ is associated with numbers $R_i \geq 0$, $\rho_i \geq 0$ and $\xi_i \in \{0, 1\}$ such that

$$R^2_{\max}(x, a) \in [R_i^2,\, (1+\delta)R_i^2] \;\;, \quad \rho^2(x, a) \in [\rho_i^2,\, (1+\delta)\rho_i^2] \;\;, \quad \xi_\gamma(x, a) = \xi_i \quad \text{for all } (x, a) \in B_i.$$

- Each B_i either satisfies $\mu(B_i) \leq \delta$ or consists of a single pair (x_i, a_i).

The numbers R_i and ρ_i will be exactly $\hat{R}(x, a)$ and $\hat{\rho}(x, a)$ from the theorem statement.

As before, we parametrize the family of reward distributions in terms of the mean reward function $\eta(x, a)$. However, now $\eta(x, a)$ is itself a random variable, which is drawn from a prior distribution. The reward function $\eta(x, a)$ will be constant on each B_i, and its value on B_i, written as $\eta(i)$, will be drawn from a scaled Bernoulli, parametrized by a prior function $\theta(i)$ as follows:

$$\eta(i) = \begin{cases} \xi_i R_i & \text{with probability } \theta(i), \\ 0 & \text{with probability } 1 - \theta(i). \end{cases} \tag{25}$$

We now set $D_\eta(r \mid x, a) = \eta(i)$ whenever $(x, a) \in B_i$. This clearly satisfies the constraints on the mean since $0 \leq \mathbb{E}[r \mid x, a] \leq R_i \leq R_{\max}(x, a)$ from the property of the partition, and also $\mathrm{Var}(r \mid x, a) = 0$ as per the setting of Theorem 4. The goal of an estimator is to take n samples generated by drawing $x \sim \lambda$, $a \mid x \sim \mu$ and $r \mid x, a \sim D_\eta$, and output an estimate $\hat{v}$ such that $\mathbb{E}_\eta[(\hat{v} - v_\eta^\pi)^2]$ is small. We recall our earlier shorthand v_η to denote the value of π under the reward distribution generated by η. For showing a lower bound on this quantity, it is clearly sufficient to pick any prior distribution governed by a parameter θ, as in Eq. (25), and lower bound the expectation $\mathbb{E}_\theta\left[\mathbb{E}_\eta[(\hat{v} - v_\eta)^2 \mid \eta]\right]$. If this expectation is large for some estimator $\hat{v}$, then there must be some realization η, which induces a large error least one function $\eta(x, a)$ which induces a large error $\mathbb{E}_\eta[(\hat{v} - v_\eta)^2 \mid \eta]$, as desired. Consequently, we focus in the proof on lower bounding the expectation $\mathbb{E}_\theta[\cdot]$. This expectation can be decomposed with the use of the inequality $a^2 \geq (a+b)^2/2 - b^2$ as follows:

$$\mathbb{E}_\theta\Big[\mathbb{E}_\eta[(\hat{v} - v_\eta)^2 \mid \eta]\Big] \geq \frac{1}{2}\mathbb{E}_\theta\Big[\mathbb{E}_\eta[(\hat{v} - \mathbb{E}_\theta[v_\eta])^2 \mid \eta]\Big] - \mathbb{E}_\theta\Big[(v_\eta - \mathbb{E}_\theta[v_\eta])^2\Big].$$

Taking the worst case over all problems in the above inequality, we obtain

$$\begin{aligned} \sup_\eta \mathbb{E}_\eta[(\hat{v} - v_\eta)^2] &\geq \sup_\theta \mathbb{E}_\theta\Big[\mathbb{E}_\eta[(\hat{v} - v_\eta)^2 \mid \eta]\Big] \\ &\geq \underbrace{\sup_\theta \frac{1}{2}\mathbb{E}_\theta\Big[\mathbb{E}_\eta[(\hat{v} - \mathbb{E}_\theta[v_\eta])^2 \mid \eta]\Big]}_{\mathcal{T}_1} - \underbrace{\sup_\theta \mathbb{E}_\theta\Big[(v_\eta - \mathbb{E}_\theta[v_\eta])^2\Big]}_{\mathcal{T}_2}. \end{aligned} \tag{26}$$

This decomposition says that the expected MSE of an estimator in estimating v_η can be related to the MSE of the same estimator in estimating the quantity $\mathbb{E}_\theta[v_\eta]$, as long as the variance of the quantity v_η under the distribution generated by θ is not too large. This is a very important observation, since we can now choose to instead study the MSE of an estimator in estimating $\mathbb{E}_\theta[v_\eta]$ as captured by $\mathcal{T}_1$. Unlike the distribution D_η which is degenerate, this problem has a non-trivial noise arising from the randomized draw of η according to θ. Thus we can use similar techniques as the proof of Theorem 3, albeit where the reward distribution is a scaled Bernoulli instead of Gaussian. For now, we focus on controlling $\mathcal{T}_1$, and $\mathcal{T}_2$ will be handled later.

In order to bound $\mathcal{T}_1$, we will consider two carefully designed choices θ_1 and θ_2 to induce two different problem instances and show that $\mathcal{T}_1$ is large for *any estimator* under one of the two parameters. In doing this, it will be convenient to use the additional shorthand $\ell_\theta(\hat{v}) = (\hat{v} - \mathbb{E}_\theta[v_\eta])^2$. Proceeding as in the proof of Theorem 3, we have

$$\begin{aligned} \mathcal{T}_1 &= \frac{1}{2}\sup_\theta \mathbb{E}_\theta\Big[\mathbb{E}_\eta[(\hat{v} - \mathbb{E}_\theta[v_\eta])^2 \mid \eta]\Big] = \frac{1}{2}\sup_\theta \mathbb{E}_\theta\Big[\mathbb{E}_\eta[\ell_\theta(\hat{v}) \mid \eta]\Big] \\ &\geq \frac{\epsilon}{2}\sup_\theta \mathbb{P}_\theta\left(\ell_\theta(\hat{v}) \geq \epsilon\right) \geq \frac{\epsilon}{2}\max_{\theta \in \theta_1, \theta_2} \mathbb{P}_\theta\left(\ell_\theta(\hat{v}) \geq \epsilon\right) \\ &\geq \frac{\epsilon}{4}\Big[\mathbb{P}_{\theta_1}\left(\ell_{\theta_1}(\hat{v}) \geq \epsilon\right) + \mathbb{P}_{\theta_2}\left(\ell_{\theta_2}(\hat{v}) \geq \epsilon\right)\Big]. \end{aligned}$$

Reduction to hypothesis testing As in the proof of Theorem 3, we now reduce the estimation problem into a hypothesis test for whether the data is generated according to the parameter θ_1 or θ_2. The arguments here are similar to the earlier proof, so we will be terser in this presentation.

As before, our hypothesis test has entire knowledge of D_η as well as θ_1 and θ_2. Consequently, we construct a test based on picking θ_1 whenever $\ell_{\theta_1}(\hat{v}) \le \ell_{\theta_2}(\hat{v})$. As before, we will ensure that $|\mathbb{E}_{\theta_1}[v_\eta] - \mathbb{E}_{\theta_2}[v_\eta]| \ge 2\sqrt{\epsilon}$ so that for any estimator $\hat{v}$, we have

$$\ell_{\theta_1}(\hat{v}) + \ell_{\theta_2}(\hat{v}) \ge 2\epsilon.$$

Under this assumption, we can similarly conclude that the error of our hypothesis test is at most

$$\mathbb{P}_{\theta_1}\left(\ell_{\theta_1}(\hat{v}) \ge \epsilon\right) + \mathbb{P}_{\theta_2}\left(\ell_{\theta_2}(\hat{v}) \ge \epsilon\right).$$

Invoking Le Cam's argument Once again, we can lower bound the error rate of our test by invoking the result of Le Cam. This requires an upper bound on the KL-divergence $D_{\mathrm{KL}}(P_{\theta_1} \| P_{\theta_2})$. The only difference from our earlier argument is that these distributions are now Bernoulli instead of Gaussian, based on the construction in Eq. (25). More formally, we have

$$\begin{aligned} D_{\mathrm{KL}}(P_{\theta_1} \| P_{\theta_2}) &= \sum_{i \in \mathcal{I}} \sum_{r \in \{0, x i_i R_i\}} \log\left(\frac{p(r; \theta_1(i))}{p(r; \theta_2(i))}\right) p(r; \theta_1(i)) \mu(B_i) \\ &= \mathbb{E}_\mu\Big[\, \xi_i D_{\mathrm{KL}}\big(\,\mathrm{Ber}(\theta_1(i)) \,\big\|\, \mathrm{Ber}(\theta_2(i))\,\big)\,\Big], \end{aligned} \tag{27}$$

where i is treated as a random variable under μ, and ξ_i is included, because the two distributions assign $r = 0$ with probability one if $\xi_i = 0$.

Picking the parameters It remains to carefully choose θ_1 and θ_2. We define $\theta_2(i) \equiv 0.5$, and let $\theta_1(i) = \theta_2(i) + \Delta_i$, where Δ_i will be chosen to satisfy certain constraints as before. Then, by Lemma 3, the KL divergence in Eq. (27) can be bounded as

$$D_{\mathrm{KL}}(P_{\theta_1} \| P_{\theta_2}) \le \frac{1}{4}\mathbb{E}_\mu\left[\xi_i \Delta_i^2\right].$$

It remains to choose Δ_i. Following a similar logic as before, we seek to find a good feasible solution of the maximization problem

$$\begin{aligned} \max_{\Delta} \quad & \epsilon \\ \text{such that} \quad & \mathbb{E}_\mu\big[\rho(x,a)\xi_i \Delta_i R_i\big] \ge 2\sqrt{\epsilon}, && (28) \\ & \frac{1}{4}\mathbb{E}_\mu\left[\xi_i \Delta_i^2\right] \le \frac{1}{n}, && (29) \\ & 0 \le \Delta_i \le 0.5. && (30) \end{aligned}$$

For some $\alpha > 0$ to be determined shortly, we set

$$\Delta_i = \min\left\{\frac{\xi_i \rho_i R_i \alpha}{\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]}, 0.5\right\}.$$

The bound constraint (30) is satisfied by construction and we set $\alpha = \sqrt{4\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]/n}$ to satisfy the constraint (29). To obtain a feasible choice of ϵ, we bound $\mathbb{E}_\mu[\rho(x,a)\xi_i\Delta_i R_i]$ as follows:

$$\begin{aligned} \mathbb{E}_\mu\big[\rho(x,a)\xi_i\Delta_i R_i\big] &\ge \mathbb{E}_\mu\big[\xi_i \rho_i \Delta_i R_i\big] \\ &\ge \mathbb{E}_\mu\left[\frac{\xi_i \rho_i^2 R_i^2 \alpha \mathbf{1}\big(\xi_i \rho_i R_i \le \mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]/2\alpha\big)}{\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]}\right] \\ &= \alpha\left(1 - \frac{\mathbb{E}_\mu\left[\xi_i \rho_i^2 R_i^2 \mathbf{1}(\xi_i \rho_i R_i > \mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]/2\alpha)\right]}{\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]}\right) \\ &=: 2\sqrt{\epsilon}. \end{aligned}$$

Collecting our arguments so far, we have established that

$$\begin{aligned}
\mathcal{T}_1 &\geq \frac{\epsilon}{4} \cdot \left(\mathbb{P}_{\theta_1} \left(\ell_{\theta_1}(\hat{v}) \geq \epsilon \right) + \mathbb{P}_{\theta_2} \left(\ell_{\theta_2}(\hat{v}) \geq \epsilon \right) \right) \\
&\geq \frac{\epsilon}{4} \cdot \frac{1}{8} e^{-n D_{\mathrm{KL}}(P_{\theta_1} \| P_{\theta_2})} \\
&\geq \frac{\epsilon}{4} \cdot \frac{1}{8e} = \frac{\epsilon}{32e} \\
&= \frac{1}{32e} \cdot \frac{\alpha^2}{4} \left(1 - \frac{\mathbb{E}_\mu \left[\xi_i \rho_i^2 R_i^2 \mathbf{1}(\xi_i \rho_i R_i > \mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]/2\alpha) \right]}{\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]} \right)^2 \\
&= \frac{\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]}{32en} \left(1 - \frac{\mathbb{E}_\mu \left[\xi_i \rho_i^2 R_i^2 \mathbf{1} \left(\xi_i \rho_i R_i > \sqrt{n \mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]/16} \right) \right]}{\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]} \right)^2 .
\end{aligned}$$

In order to complete the proof, we need to further upper bound $\mathcal{T}_2$ in the decomposition (26).

Bounding $\mathcal{T}_2$ We need to bound the supremum over all priors θ. Consider an arbitrary prior θ and assume that η is drawn according to Eq. (25). To bound $\mathbb{E}_\theta \left[(v_\eta - \mathbb{E}_\theta[v_\eta])^2 \right]$, we view $(v_\eta - \mathbb{E}_\theta[v_\eta])^2$ as a random variable under θ and bound it using Hoeffding's inequality.

We begin by bounding its range. From the definition of η and v_η,

$$0 \leq v_\eta \leq \mathbb{E}_\pi[\xi_i R_i] = \mathbb{E}_\mu[\rho(x,a) \xi_i R_i] \leq (1+\delta)^{1/2} \mathbb{E}_\mu[\xi_i \rho_i R_i] \ ,$$

so also $0 \leq \mathbb{E}_\theta[v_\eta] \leq (1+\delta)^{1/2} \mathbb{E}_\mu[\xi_i \rho_i R_i]$. Hence, $|v_\eta - \mathbb{E}_\theta[v_\eta]| \leq (1+\delta)^{1/2} \mathbb{E}_\mu[\xi_i \rho_i R_i]$, and we obtain the bound

$$(v_\eta - \mathbb{E}_\theta[v_\eta])^2 \leq (1+\delta)(\mathbb{E}_\mu[\xi_i \rho_i R_i])^2 \leq (1+\delta) \mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]. \tag{31}$$

The proof proceeds by applying Hoeffding's inequality to control the probability that $(v_\eta - \mathbb{E}_\theta[v_\eta])^2 \geq t^2$ for a suitable t. Then we can, with high probability, use the bound $(v_\eta - \mathbb{E}_\theta[v_\eta])^2 \geq t^2$, and with the remaining small probability apply the bound of Eq. (31).

To apply Hoeffding's inequality, we write v_η explicitly as

$$v_\eta = \sum_{i \in \mathcal{I}} \mu(B_i) \rho_i' \eta_i =: \sum_{i \in \mathcal{I}} Y_i$$

where $\rho_i' := \mathbb{E}_\mu[\rho(x,a) \mid (x,a) \in B_i]$. Thus, v_η can be written as a sum of countably many independent variables, but we can only apply Hoeffding's inequality to their finite subset. Note that the variables Y_i are non-negative and upper-bounded by a summable series, namely $Y_i \leq \mu(B_i) \rho_i' R_i$, where the summability follows because $\mathbb{E}_\mu[\rho R_{\max}] \leq 1 + \mathbb{E}_\mu[\rho^2 R_{\max}^2] < \infty$. This means that for any $\delta_0 > 0$, we can choose a finite set $\mathcal{I}_0$ such that $\sum_{i \notin \mathcal{I}_0} Y_i \leq \delta_0$. We will determine the sufficiently small value of δ_0 later; for now, consider the corresponding set $\mathcal{I}_0$ and define an auxiliary variable

$$v_\eta' := \sum_{i \in \mathcal{I}_0} Y_i \ ,$$

which by construction satisfies $v_\eta' \leq v_\eta \leq v_\eta' + \delta_0$. Note that the summands Y_i can be bounded as

$$0 \leq Y_i \leq \xi_i \rho_i' R_i \mu(B_i) \leq \xi_i (1+\delta)^{1/2} \rho_i R_i \sqrt{\mu(B_i)} \sqrt{\gamma'}$$

because $\rho_i' \leq (1+\delta)^{1/2} \rho_i$ and $\xi_i \mu(B_i) \leq \xi_i \sqrt{\mu(B_i)} \sqrt{\gamma'}$, because $\xi_i = 0$ whenever $\mu(B_i) > \max\{\gamma, \delta\} = \gamma'$. By Hoeffding's inequality, we thus have

$$\begin{aligned}
\mathbb{P}(|v_\eta' - \mathbb{E}_\theta v_\eta'| \geq t) &\leq 2 \exp \left\{ - \frac{2t^2}{(1+\delta) \sum_{i \in \mathcal{I}_0} \xi_i \rho_i^2 R_i^2 \mu(B_i) \gamma'} \right\} \\
&\leq 2 \exp \left\{ - \frac{2t^2}{(1+\delta) \gamma' \mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]} \right\}.
\end{aligned}$$

Now take $t = \sqrt{\gamma' \log(4/\gamma')(1+\delta)\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]/2}$ in the above bound, which yields

$$\mathbb{P}\Big[(v'_\eta - \mathbb{E}_\theta v'_\eta)^2 \geq t^2\Big] = \mathbb{P}\Big[|v'_\eta - \mathbb{E}_\theta v'_\eta| \geq t\Big] \leq \frac{\gamma'}{2} \ .$$

Now, we can go back to analyzing v_η. We set δ_0 sufficiently small, so $t + \delta_0 \leq \sqrt{\gamma' \log(5/\gamma')(1+\delta)\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]/2}$. Thus, using Eq. (31), we have

$$\begin{aligned}
\mathbb{E}_\theta\left[(v_\eta - \mathbb{E}_\theta v_\eta)^2\right] &\leq (t+\delta_0)^2 \cdot \mathbb{P}\Big[(v_\eta - \mathbb{E}_\theta v_\eta)^2 < (t+\delta_0)^2\Big] \\
&\qquad + (1+\delta)\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2] \cdot \mathbb{P}\Big[(v_\eta - \mathbb{E}_\theta v_\eta)^2 \geq (t+\delta_0)^2\Big] \\
&\leq (t+\delta_0)^2 + (1+\delta)\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2] \cdot \mathbb{P}\Big[(v'_\eta - \mathbb{E}_\theta v'_\eta)^2 \geq t^2\Big] \\
&= \frac{\gamma' \log(5/\gamma')(1+\delta)\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2]}{2} + (1+\delta)\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2] \cdot \frac{\gamma'}{2} \\
&\leq \gamma' \log(5/\gamma')(1+\delta)\mathbb{E}_\mu[\xi_i \rho_i^2 R_i^2] \ .
\end{aligned}$$

Combining this bound with the bound on $\mathcal{T}_1$ yields the theorem.

Lemma 1. *Let $\mathcal{Z} \coloneqq \mathcal{X} \times \mathcal{A}$ be a subset of $\mathbb{R}^d$, let μ be a probability measure on $\mathcal{Z}$ and $R_{\max}$ and ρ be non-negative measurable functions on $\mathcal{Z}$. Given $\gamma \in [0,1]$, define a random variable $\xi_\gamma(z) \coloneqq \mathbf{1}(\mu(z) \leq \gamma)$. Then for any $\delta \in (0,1]$, there exists a countable index set $\mathcal{I}$ and disjoint sets $B_i \subseteq \mathcal{Z}$ alongside non-negative reals R_i, ρ_i and $\xi_i \in \{0,1\}$ such that the following conditions hold:*

- *Sets B_i form a partition of $\mathcal{Z}$, i.e., $\mathcal{Z} = \uplus_{i \in \mathcal{I}} B_i$.*
- *Reals R_i and ρ_i approximate $R_{\max}$ and ρ, and ξ_i equals ξ_γ as follows:*

$$R_{\max}^2(z) \in [R_i^2,\ (1+\delta)R_i^2] \ , \quad \rho^2(z) \in [\rho_i^2,\ (1+\delta)\rho_i^2] \ , \quad \xi_\gamma(z) = \xi_i \quad \textit{for all } z \in B_i.$$

- *Each set B_i either satisfies $\mu(B_i) \leq \delta$ or consists of a single $z \in \mathcal{Z}$.*

Proof. Let $\mathcal{Z} \coloneqq \mathcal{X} \times \mathcal{A}$. We begin our construction by separating out atoms, i.e., the elements $z \in \mathcal{Z}$ such that $\mu(z) > 0$. Specifically, we write $\mathcal{Z} = \mathcal{Z}^{\mathrm{na}} \uplus \mathcal{Z}^{\mathrm{a}}$ where $\mathcal{Z}^{\mathrm{a}}$ consists of atoms and $\mathcal{Z}^{\mathrm{na}}$ of all non-atoms. The set $\mathcal{Z}^{\mathrm{a}}$ is either finite or countably infinite, so $\mathcal{Z}^{\mathrm{na}}$ is measurable.

By a theorem of Sierpiński (1922), since μ does not have any atoms on $\mathcal{Z}^{\mathrm{na}}$, it must be continuous on $\mathcal{Z}^{\mathrm{na}}$ in the sense that if A is a measurable subset of $\mathcal{Z}^{\mathrm{na}}$ with $\mu(A) = a$ then for any $b \in [0,a]$, there exists a measurable set $B \subseteq A$ such that $\mu(B) = b$. This means that we can decompose $\mathcal{Z}^{\mathrm{na}}$ into $N \coloneqq \lceil 1/\delta \rceil$ sets $\mathcal{Z}_1^{\mathrm{na}}, \mathcal{Z}_2^{\mathrm{na}}, \ldots, \mathcal{Z}_N^{\mathrm{na}}$ such that each has a measure at most δ and $\mathcal{Z}^{\mathrm{na}} = \biguplus_{j=1}^N \mathcal{Z}_j^{\mathrm{na}}$.

We next ensure the approximation properties for $R_{\max}$ and ρ. We begin by a countable decomposition of non-negative reals. We consider the countable index set $\mathcal{J} \coloneqq \mathbb{Z} \cup \{-\infty\}$ and define the sequence $a_j \coloneqq (1+\delta)^{j/2}$, for $j \in \mathbb{Z}$. Positive reals can then be decomposed into the following intervals indexed by $\mathcal{J}$:

$$I_{-\infty} \coloneqq \{0\} \ , \qquad I_j \coloneqq (a_j, a_{j+1}] \quad \text{for } j \in \mathbb{Z}.$$

It will also be convenient to set $a_{-\infty} \coloneqq 0$. Thus, the construction of I_j guarantees that for all $j \in \mathcal{J}$ and all $t \in I_j$ we have $a_j^2 \leq t^2 \leq (1+\delta)a_j^2$.

The desired partition, with the index set $\mathcal{I} = \mathcal{Z}^{\mathrm{a}} \ \cup \ [N] \times \mathcal{J}^2$, is as follows:

$$\begin{aligned}
&\text{for } i = z \in \mathcal{Z}^{\mathrm{a}}: && B_i \coloneqq \{z\},\ R_i \coloneqq R_{\max}(z),\ \rho_i \coloneqq \rho(z),\ \xi_i \coloneqq \xi_\gamma(z); \\
&\text{for } i = (j, j_R, j_\rho) \in [N] \times \mathcal{J}^2: && B_i \coloneqq \mathcal{Z}_j^{\mathrm{na}} \cap R_{\max}^{-1}(I_{j_R}) \cap \rho^{-1}(I_{j_\rho}), \\
& && R_i \coloneqq a_{j_R},\ \rho_i \coloneqq a_{j_\rho},\ \xi_i \coloneqq 1.
\end{aligned}$$

□

B. Proof of Theorem 2

Let $A_x \coloneqq \{a \in \mathcal{A} :\ \rho(x,a) \le \tau\}$. For brevity, we write $A_i \coloneqq A_{x_i}$. We decompose the mean squared error into the squared bias and variance and control each term separately,

$$\mathrm{MSE}(\hat{v}_{\textsc{switch}}) = \left|\mathbb{E}[\hat{v}_{\textsc{switch}}] - v^\pi\right|^2 + \mathrm{Var}[\hat{v}_{\textsc{switch}}].$$

We first calculate the bias. Note that bias is incurred only in the terms that fall in A_x^c, so

$$\begin{aligned}
\mathbb{E}[\hat{v}_{\textsc{switch}}] - v^\pi &= \mathbb{E}\left[\sum_{a\in A_x^c} \hat{r}(x,a)\pi(a|x)\right] - \mathbb{E}\left[\sum_{a\in A_x^c} \mathbb{E}[r|x,a]\,\pi(a|x)\right] \\
&= \mathbb{E}_\pi\Big[\big(\hat{r}(x,a) - \mathbb{E}[r|x,a]\big)\,\mathbf{1}(a \in A_x^c)\Big] \\
&= \mathbb{E}_\pi\big[\,\epsilon(x,a)\,\mathbf{1}(\rho > \tau)\,\big]
\end{aligned}$$

where we recall that $\epsilon(x,a) = \hat{r}(x,a) - \mathbb{E}[r|x,a]$.

Next we upper bound the variance. Note that the variance contributions from the IPS part and the DM part are not independent, since the indicators $\rho(x_i,a) > \tau$ and $\rho(x_i,a) \le \tau$ are mutually exclusive. To simplify the analysis, we use the following inequality that holds for any random variable X and Y:

$$\mathrm{Var}(X+Y) \le 2\mathrm{Var}(X) + 2\mathrm{Var}(Y).$$

This allows us to calculate the variance of each part separately.

$$\begin{aligned}
\mathrm{Var}[\hat{v}_{\textsc{switch}}] &\le 2\,\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n [r_i\rho_i\mathbf{1}(a_i \in A_i)]\right] + 2\,\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n\sum_{a\in\mathcal{A}} \hat{r}(x_i,a)\pi(a|x_i)\mathbf{1}(a\in A_i^c)\right] \\
&= \frac{2}{n}\,\mathrm{Var}_\mu\big[r\rho\mathbf{1}(a\in A_x)\big] + \frac{2}{n}\,\mathrm{Var}\left[\sum_{a\in A_x^c}\hat{r}(x,a)\pi(a|x)\right] \\
&= \frac{2}{n}\,\mathbb{E}_\mu\mathrm{Var}\big[r\rho\mathbf{1}(a\in A_x)\,\big|\,x,a\big] + \frac{2}{n}\,\mathrm{Var}_\mu\mathbb{E}\big[r\rho\mathbf{1}(a\in A_x)\,\big|\,x,a\big] + \frac{2}{n}\,\mathrm{Var}\left[\sum_{a\in A_x^c}\hat{r}(x,a)\pi(a|x)\right] \\
&\le \frac{2}{n}\,\mathbb{E}_\mu\mathrm{Var}\big[r\rho\mathbf{1}(a\in A_x)\,\big|\,x,a\big] + \frac{2}{n}\,\mathbb{E}_\mu\Big[\mathbb{E}[r\rho\mathbf{1}(a\in A_x)\mid x,a]^2\Big] + \frac{2}{n}\,\mathbb{E}\left[\Big(\sum_{a\in A_x^c}\hat{r}(x,a)\pi(a|x)\Big)^2\right] \\
&\le \frac{2}{n}\,\mathbb{E}_\mu\big[\sigma^2\rho^2\mathbf{1}(a\in A_x)\big] + \frac{2}{n}\,\mathbb{E}_\mu\big[R_{\max}^2\rho^2\mathbf{1}(a\in A_x)\big] + \frac{2}{n}\,\mathbb{E}\left[\Big(\sum_{a\in A_x^c}\hat{r}(x,a)\pi(a|x)\Big)^2\right].
\end{aligned}$$

To complete the proof, note that the last term is further upper bounded using Jensen's inequality as

$$\begin{aligned}
\mathbb{E}\left[\Big(\sum_{a\in A_x^c}\hat{r}(x,a)\pi(a|x)\Big)^2\right] &= \mathbb{E}\left[\Big(\sum_{a\in A_x^c}\pi(a|x)\Big)^2\Big(\sum_{a\in A_x^c}\frac{\hat{r}(x,a)\pi(a|x)}{\sum_{a\in A_x^c}\pi(a|x)}\Big)^2\right] \\
&\le \mathbb{E}\left[\Big(\sum_{a\in A_x^c}\pi(a|x)\Big)\Big(\sum_{a\in A_x^c}\hat{r}(x,a)^2\pi(a|x)\Big)\right] \\
&\le \mathbb{E}_\pi\big[R_{\max}^2\mathbf{1}(\rho>\tau)\big],
\end{aligned}$$

where the final inequality uses $\sum_{a\in A_x^c}\pi(a|x) \le 1$ and $\hat{r}(x,a) \in [0, R_{\max}(x,a)]$ almost surely.

Combining the bias and variance bounds, we get the stated MSE upper bound. □

C. Utility Lemmas

Lemma 2 (Hoeffding, 1963, Theorem 2)**.** *Let $X_i \in [a_i, b_i]$ and $X_1, ..., X_n$ are drawn independently. Then the empirical mean $\bar{X} = \frac{1}{n}(X_1 + ... + X_n)$ obeys*

$$\mathbb{P}(|\bar{X} - \mathbb{E}[\bar{X}]| \geq t) \leq 2e^{-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}.$$

Lemma 3 (Bernoulli KL-divergence)**.** *For $0 < p, q < 1$, we have*

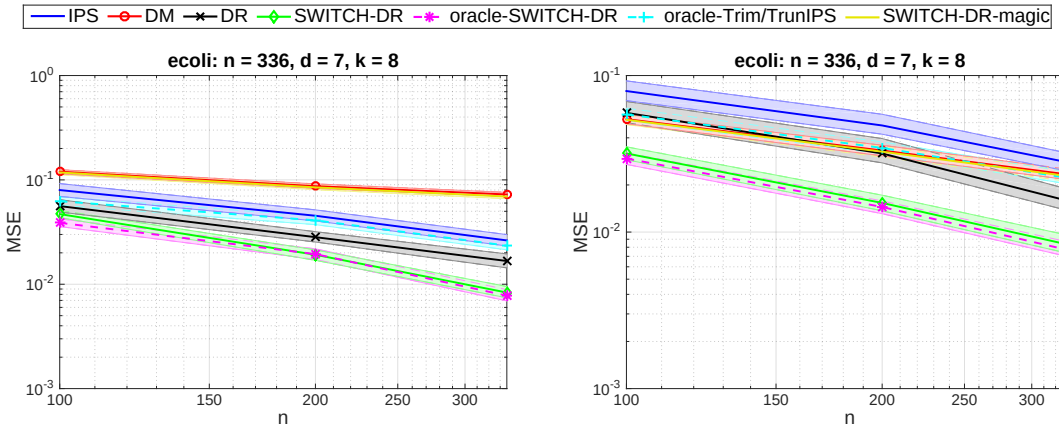
$$D_{\mathrm{KL}}(\mathrm{Ber}(p) \| \mathrm{Ber}(q)) \leq (p-q)^2 (\frac{1}{q} + \frac{1}{1-q}).$$

Proof.

$$\begin{aligned} D_{\mathrm{KL}}(\mathrm{Ber}(p) \| \mathrm{Ber}(q)) &= p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right) \\ &\leq p\frac{p-q}{q} + (1-p)\frac{q-p}{1-q} = \frac{(p-q)^2}{q} + (p-q) + \frac{(p-q)^2}{1-q} + (q-p) \\ &= (p-q)^2\left(\frac{1}{q} + \frac{1}{1-q}\right). \end{aligned}$$

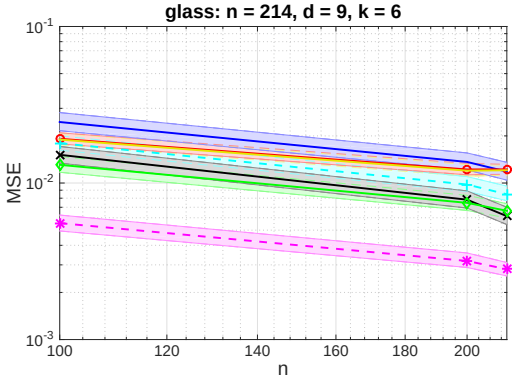
□

D. Additional Figures from the Experiments

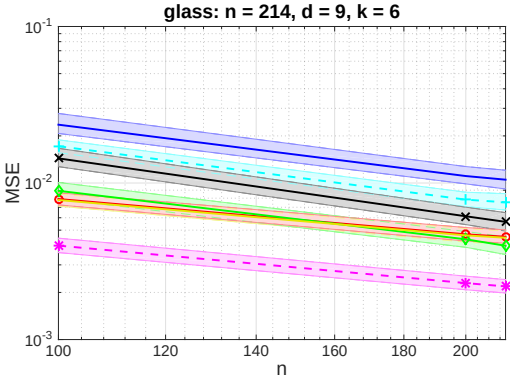


(a) ecoli / deterministic reward

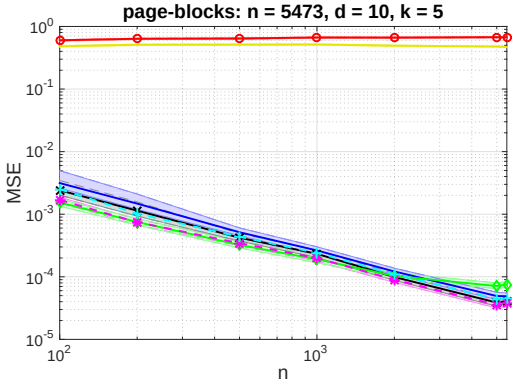
(b) ecoli / noisy reward



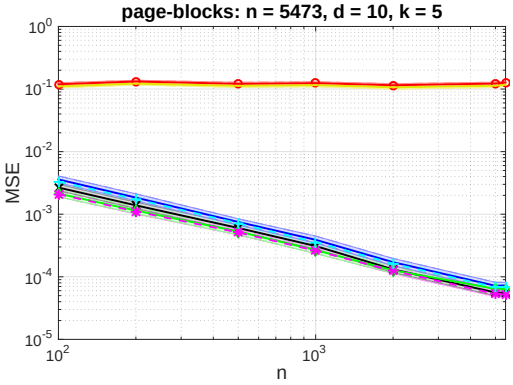
(c) glass / deterministic reward



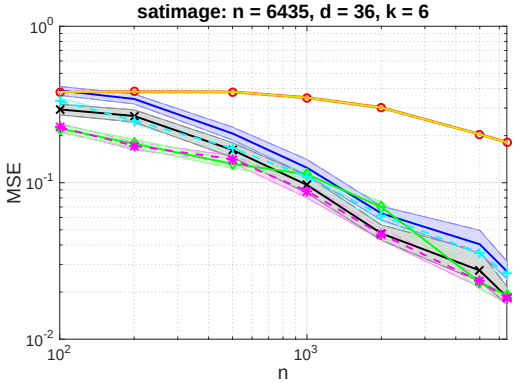
(d) glass / noisy reward



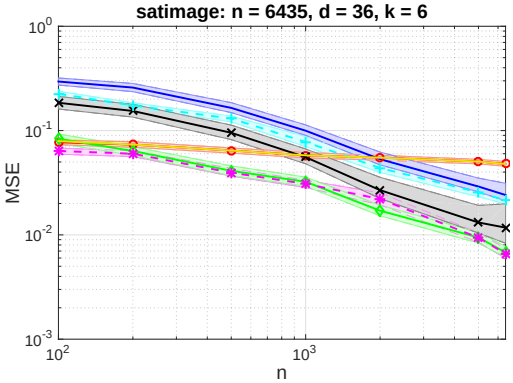
(e) page-blocks / deterministic reward



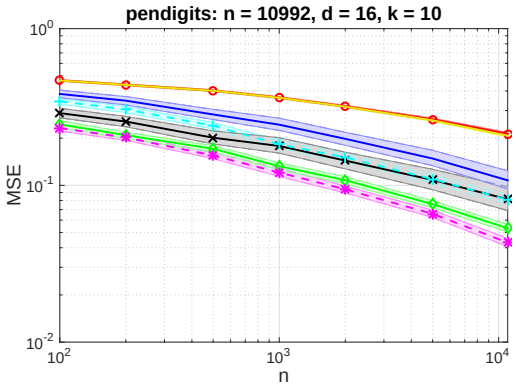
(f) page-blocks / noisy reward



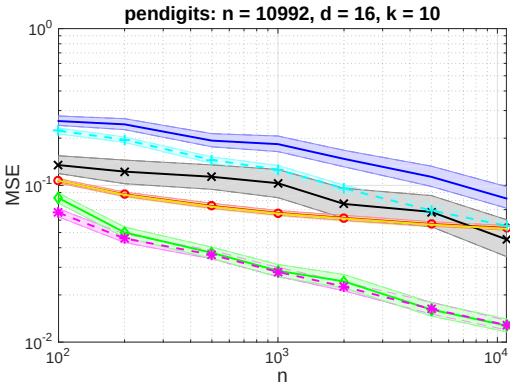
(g) satimage / deterministic reward



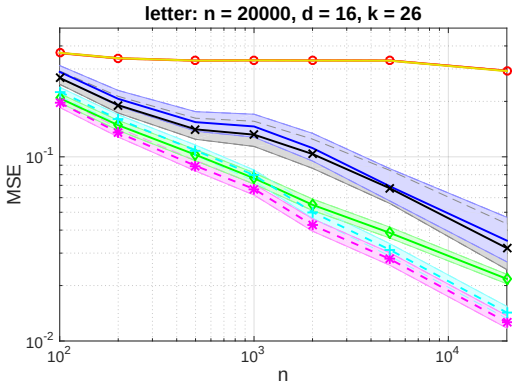
(h) satimage / noisy reward



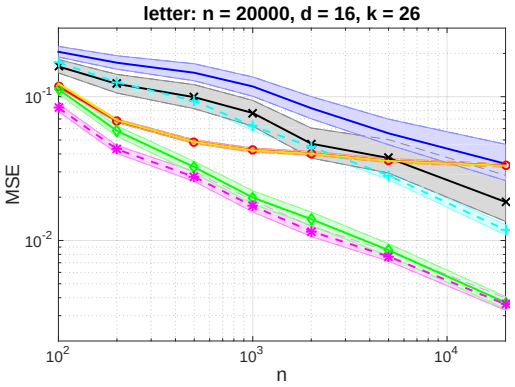
(a) pendigits / deterministic reward



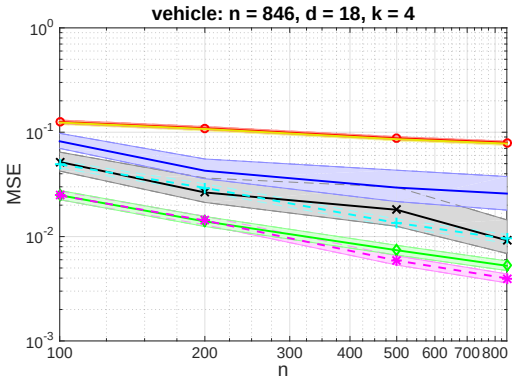
(b) pendigits / noisy reward



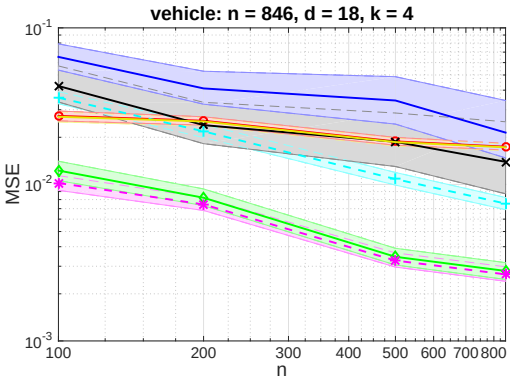
(c) letter / deterministic reward



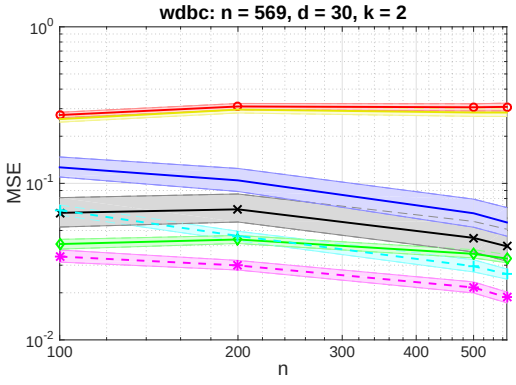
(d) letter / noisy reward



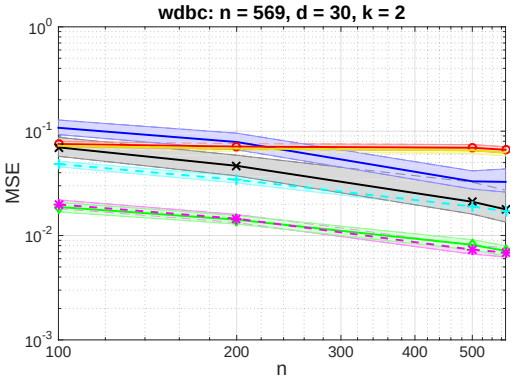
(e) vehicle / deterministic reward



(f) vehicle / noisy reward



(g) wdbc / deterministic reward



(h) wdbc / noisy reward